

Supplementary Information for

Genome-wide Identification of Zero Nucleotide Recursive Splicing in *Drosophila*

Michael Duff^{1*}, Sara Olson^{1*}, Xintao Wei^{1*}, Ahmad Osman¹, Alex Plocik¹, Mohan Bolisetty¹, Susan Celniker², and Brenton R. Graveley^{1,3}

¹Department of Genetics and Developmental Biology, Institute for Systems Genomics, University of Connecticut Health Center, Farmington, Connecticut 06030, USA

²Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

³Corresponding author: graveley@uchc.edu

*These authors contributed equally to this work

Index to Supplementary Results and Methods

Supplementary Methods	3
RNA Collections	3
RNA Sequencing	3
Alignments	3
Computational identification of ratchet points	3
Parsing TopHat alignments to identify ratchet points	3
de novo identification of ratchet points	4
Verification of potential ratchet points	5
Validation of recursive splice sites by RT-PCR	6
Validation of recursive splice sites by Cross-species RNA Sequencing	7
Identification of recursive lariat introns and branchpoint analysis	7
Comparison of gene expression and recursive splicing	8
Analysis of chromatin marks	9
Supplementary Figures	11
Supplementary Figure 1	
Characteristics of recursive splicing from total RNA-Seq data	11
Supplementary Figure 2	
Two approaches for identifying recursive splice sites	12
Supplementary Figure 3	
RT-PCR validations of recursive splicing events	13
Supplementary Figure 4	
Number of mapped reads per sample used for gene expression analysis	14
Supplementary Figure 5	
Chromatin marks associated with recursive splice sites	15
Supplementary Tables	17
Supplementary Table 1	
Summary of <i>D. melanogaster</i> RNA Sequencing Samples Generated for these studies	17
Supplementary Table 2	
Summary information for 197 <i>D. melanogaster</i> ratchet points	17
Supplementary Table 3	
Comparison of ratchet points identified in this study and in Burnette <i>et al.</i>	17
Supplementary Table 4.	
Summary of recursive intron lariats identified	18
Supplementary Table 5.	
Summary of Total RNA-Seq Data from Related <i>Drosophila</i> Species	18
Supplementary Table 6.	
Summary of ratchet points experimentally validated in other <i>Drosophila</i> species	18
Supplementary Table 7.	
Details of ratchet points experimentally validated in other <i>Drosophila</i> species	18
Supplementary Table 8.	
GO analysis of recursively spliced genes	18
Supplementary References	19

SUPPLEMENTARY METHODS

RNA collections. The *D. melanogaster* RNA samples used for this study were previously described^{1,2}. RNA was isolated from *D. simulans*, *D. sechellia*, *D. yakuba*, *D. pseudoobscura*, and *D. virilis* mixed adults using Trizol.

RNA sequencing. Total RNA-Seq libraries were prepared using Illumina TruSeq Stranded Total RNA Sample Prep Kits as described by the manufacturer. Libraries were quantitated by analysis on an Agilent Bioanalyzer or TapeStation and sequenced on an Illumina HiSeq2000 to generate paired-end 100 bp reads.

Alignments. Total RNA strand-specific paired-end sequence data from *D. melanogaster* was aligned to the *D. melanogaster* genome (Release 5, dm3) lacking chromosome U extra, guided by the modENCODE annotation MDv1^(ref. 1) using TopHat³ version 1.4.1 with the following settings: -p 8 -z0 -a 6 -m 2 --min-intron-length 28 -l 200000 -g 1 --library-type fr-firststrand -x 60 -n 2. The *D. simulans*, *D. sechellia*, *D. yakuba*, *D. pseudoobscura*, and *D. virilis* RNA-Seq datasets were aligned to the *D. simulans* (droSim1), *D. sechellia* (droSec1), *D. yakuba* (droYak2), *D. pseudoobscura* (dp4), and *D. virilis* (droVir3) genomes, respectively, using the same method and parameters, but without a reference transcriptome annotation.

Computational identification of ratchet points.

Parsing TopHat alignments to identify ratchet points. We identified sets of novel splice junctions from the TopHat alignments which share the same 5' splice site. Ratchet

point junctions were kept from a sample if the 3' splice site of the junction is an AG|GT, the 3' splice site was unannotated in the previous modENCODE annotation¹, and the distance to the previous splice junction and the next splice junction is longer than 2 kbp.

de novo identification of ratchet points. We generated a set of potential ratchet point junctions by joining 95 nt of each exon to the 95 nt downstream of every unannotated 3' splice site (AG|GT) in the downstream intron. We aligned reads 1 and 2 of the total RNA reads independently to the database of all possible ratchet point junctions using Bowtie⁴ version 0.12.7 with the following options: -v 2 -k 5 -M 5 --best. As the paired-end reads were aligned separately, post-processing was used to enforce constraints on gene-strand and alignment-strand that are a consequence of the stranded protocol. For each potential ratchet-point junction, we tabulated the coordinates of the genomic regions that comprise the ratchet-point junction sequences, the intron(s) and gene(s) the ratchet-site is derived from, the number of alignments to the ratchet-point junction, the average number of mismatches per alignment, detailed offset and mismatch information, the alignment offset entropy¹, and the number of distinct offsets for only perfect alignments with ≥ 8 nt overhang. This latter parameter is intended to be a robust and conservative measure of alignment diversity. Ratchet points contained in introns that overlap no other distinct introns or any exons, were filtered to require ≥ 3 perfect alignments to 3 distinct offsets & $\text{overhang} \geq 8$ (or 2 perfect alignments to 2 distinct offsets & $\text{overhang} \geq 8$ AND ≥ 10 general alignments with ≤ 2 mismatch & $\text{overhang} \geq 5$). Ratchet points in introns that do overlap other introns or exons, were

filtered using slightly more-restrictive criteria and required at least 5 zero-mismatch ratchet-junction alignments with distinct offsets).

Verification of potential ratchet points. We next compared the lists of potential ratchet points individually identified by analysis of the TopHat alignments and by alignment to all potential ratchet point junctions. This resulted in a list of 356 potential ratchet points that were then individually examined on the genome browser to verify their identity. To facilitate this analysis we merged the bedGraphs from all of the TopHat alignments into one positive- and one negative-strand-specific bedGraph file. For each intron containing potential ratchet points, we calculated a robust linear regression of the read density of each segment of the recursive introns from the merged bedGraph profile using the robustfit feature of MATLAB. This required masking out repeatMasker regions and overlapping annotated features, both of which confound the regression process, and calculating the robust regression lines based on the remaining unmasked portions of the recursive segment. We used MATLAB to generate a “flip-book” of browser-like images of each potential recursive intron with bedGraph and local robust regression plots superimposed to aid in the manual inspection of each ratchet point for verification.

The merged bedGraphs were loaded into the genome browser along with tracks of all 356 potential ratchet point splice junctions. In addition, we loaded the FlyBase 5.45 annotation, which was the most recent version of FlyBase at the time of this analysis, as well as the most recent modENCODE annotation (MDv3)² to identify ratchet points that corresponded to exons identified more recently than the modENCODE annotation (MDv1)¹ used to seed the alignments. Finally, we loaded in the

modENCODE CAGE data² to assess whether any potential ratchet points corresponded to previously unannotated promoters, which could also give rise to a sawtooth pattern of RNA-Seq read density. Potential ratchet points were removed if the 3' ratchet point junction corresponded to an annotated exon, or if the sawtooth pattern of read density was not apparent on the browser or from the local robust regression plots. During the course of this manual inspection we identified two ratchet points in *luna* and *mbi* that were not identified in this computational analysis. Both of these were present in introns that contained other computationally-identified ratchet points. We identified these based on their strong pattern of sawtooth read density in both the browser and the local robust regression plots and the fact that they had conserved AG/GT sequences at the ratchet point junctions. These were missed computationally because they did not pass the stringent filters used. Both of these were experimentally validated and included in the analysis of all ratchet points. In total, the final list of ratchet points consists of 197 ratchet points (Supplementary Table 2).

Validation of recursive splice sites by RT-PCR. 500 ng of total RNA isolated from S2 cells with Trizol was used to synthesize cDNA using SuperScript® II Reverse Transcriptase (RT) kits according to the manufacturers protocol. PCR amplification was performed using Phusion® High-Fidelity DNA Polymerase (NEB) according to the manufacturer's instructions using specific primers with the following amplification program: 98°C for 3 min, followed by 40 cycles 98°C for 10 sec, 55°C for 30 sec, and 72°C for 20 sec. Ratchet points were first confirmed by gel electrophoresis followed by Sanger sequencing.

Validation of recursive splice sites by cross-species RNA sequencing. The coordinates of the *D. melanogaster* ratchet points, and 50 nt on either side, were lifted over to the *D. simulans* (droSim1), *D. sechellia* (droSec1), *D. yakuba* (droYak2), *D. pseudoobscura* (dp4), and *D. virilis* (droVir3) genomes using the UCSC liftover tool on galaxy. The TopHat alignments of the *D. simulans*, *D. sechellia*, *D. yakuba*, *D. pseudoobscura*, and *D. virilis* were searched for splice junction reads whose 3' end mapped within the lifted over coordinates and which had an AG/GT at the recursive junction.

Identification of recursive lariat introns and branchpoint analysis. To generate potential junctions between the 5' splice site and branchpoints of intron lariats, we used a custom perl script to fused the last 94 nt of an intron to the first 94 nt of the intron. We used 94 nt of each portion of the intron to enforce a minimum of a 6 nt overhang when aligning 100 nt reads. Because the precise location of the branchpoints are not know ahead of time, and because a previous study in human has shown that most known branchpoints occur between 18 and 35 nt upstream of the 3' splice site⁵, we generated 100 possible lariat junctions by sliding a window of the 3' end of the intron in the 5' direction at 1 nt intervals, and fusing each to the first 94 nt of the intron. We generated the potential lariat junction databases for each segment of all recursive introns, as well as all possible permutations of these segments. For example, for an intron with two ratchet points, we generated potential lariat junctions for the first, second and third recursive segments as well as the introns from the first 5' splice site to the second

ratchet point (segments 1 and 2), from the 5' splice site of the first ratchet point to the last 3' splice site (segments 2 and 3), and from the first 5' splice site to the last 3' splice site (segments 1, 2 and 3).

Bowtie⁴ version 0.12.7 was used to generate an index of the potential recursive lariat junctions and to align all of the total RNA-Seq reads, where each mate pair was aligned separately, with the following parameters: -v 2 -p 8 --all --quiet. In total, 72,712 alignments were reported. These were then filtered for reads that mapped uniquely to the lariat junction database, then sorted by the number of nucleotides overlapping the lariat junction. Randomly selected reads with various extents of overlap were aligned to the *D. melanogaster* genome using BLAT to determine whether or not they mapped elsewhere in the genome. From this we determined that all reads that overlapped the lariat junction with fewer than 14 nt also aligned elsewhere in the genome. We therefore used BLAT to align all reads reported from the Bowtie alignment and discarded those which mapped elsewhere resulting in a total of 46 reads. We identified the approximate locations of the branch points based on the coordinates of the 94 nt segment of the 3' end of the intron that the reads aligned to.

Comparison of gene expression and recursive splicing. The total number of reads mapping to each ratchet point junction was tabulated for each library and then summed across biological and/or technical replicates for each biological sample. The recursive index was calculated as the number of ratchet point junction read per mapped reads per billion reads ((ratchet point junction reads/mapped reads) x 1,000,000,000). The RNA-seq data generated for this study is from total RNA, a mixture of mRNA, nascent RNA,

and pre-mRNA. Due to the complexity of RNA types and intronic reads it is difficult to accurately quantitate gene expression levels when using most existing software. We therefore used the expression values calculated from the corresponding poly(A)⁺ RNA-Seq data that was previously generated^{1,2} from the same RNA samples. Comparisons of the recursive index and gene expression levels were performed using R (<http://www.r-project.org>).

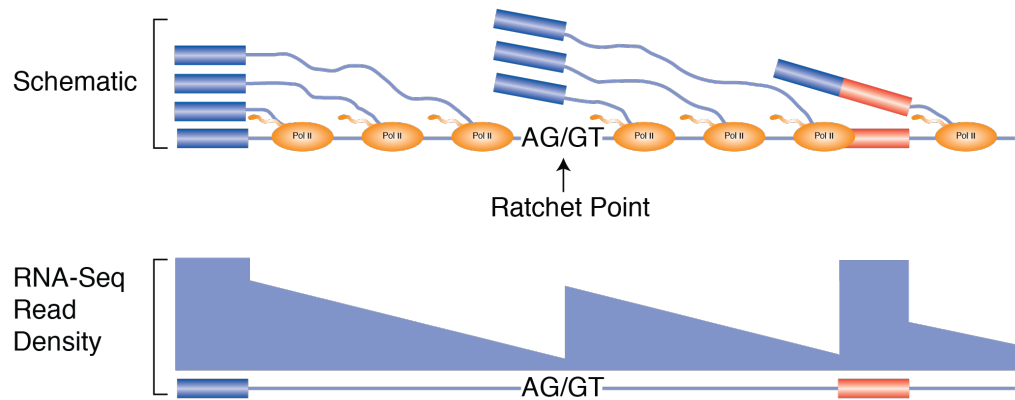
Analysis of chromatin marks. Visualizations of chromatin marks at recursively spliced genes were generated using custom R scripts (<http://www.r-project.org>). We obtained ChIP-seq scores (http://encode-x.med.harvard.edu/data_sets/chromatin/) and Affymetrix tiling array gene expression scores (<http://intermine.modencode.org/>) generated from L3 larvae via the modENCODE projects. For each feature of gene architecture illustrated, mean ChIP-seq scores were calculated for non-overlapping bins of 200 bp in length.

As expected, transcription-associated marks were specific to genes actively transcribed in larvae (Supplementary Fig. 5b). At these active genes, we observed low levels of H3K4me3 near recursive splice sites compared to first exons (Supplementary Fig. 5b), which suggests that the saw-tooth patterns observed by total RNA-seq were not due to cryptic transcription initiation or unannotated promoters, but rather, co-transcriptional splicing. We also observed lower levels of H3K36me3 near recursive splice sites compared to downstream exons (Supplementary Fig. 5b). Since the degree of H3K36me3 has been shown to increase with each internal exon in humans⁶, the low levels of H3K36me3 seen at recursive splice may reflect the fact that recursive splices are typically located in 5' introns, and thus, preceded by few internal exons

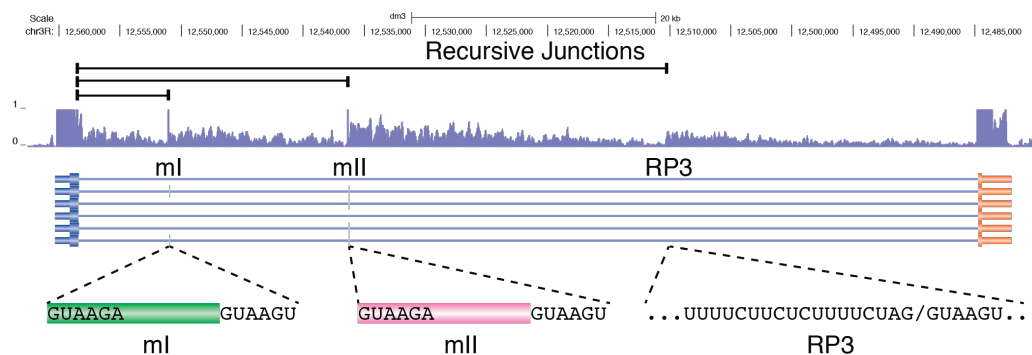
(Supplementary Fig. 5c). Indeed, recursive splice sites were associated with high levels of H3K79me2 exons, which is typical of long 5' introns in humans⁷.

Supplementary Figures

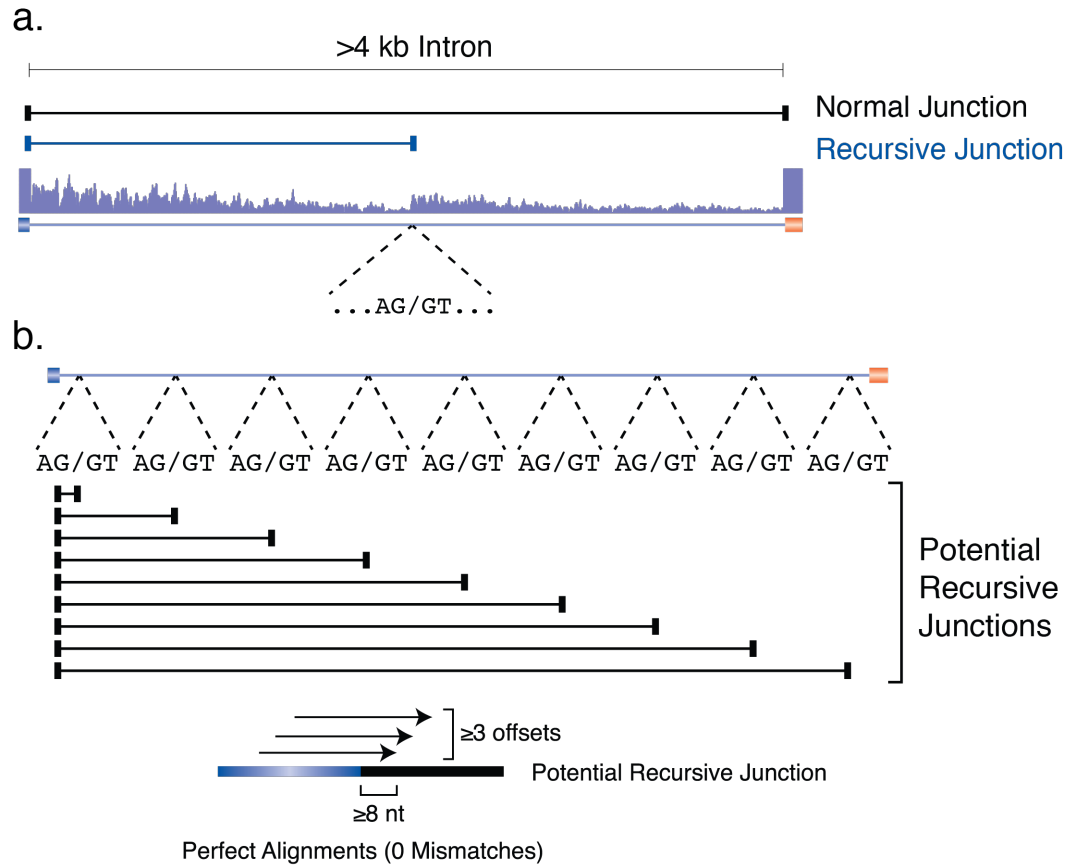
a.



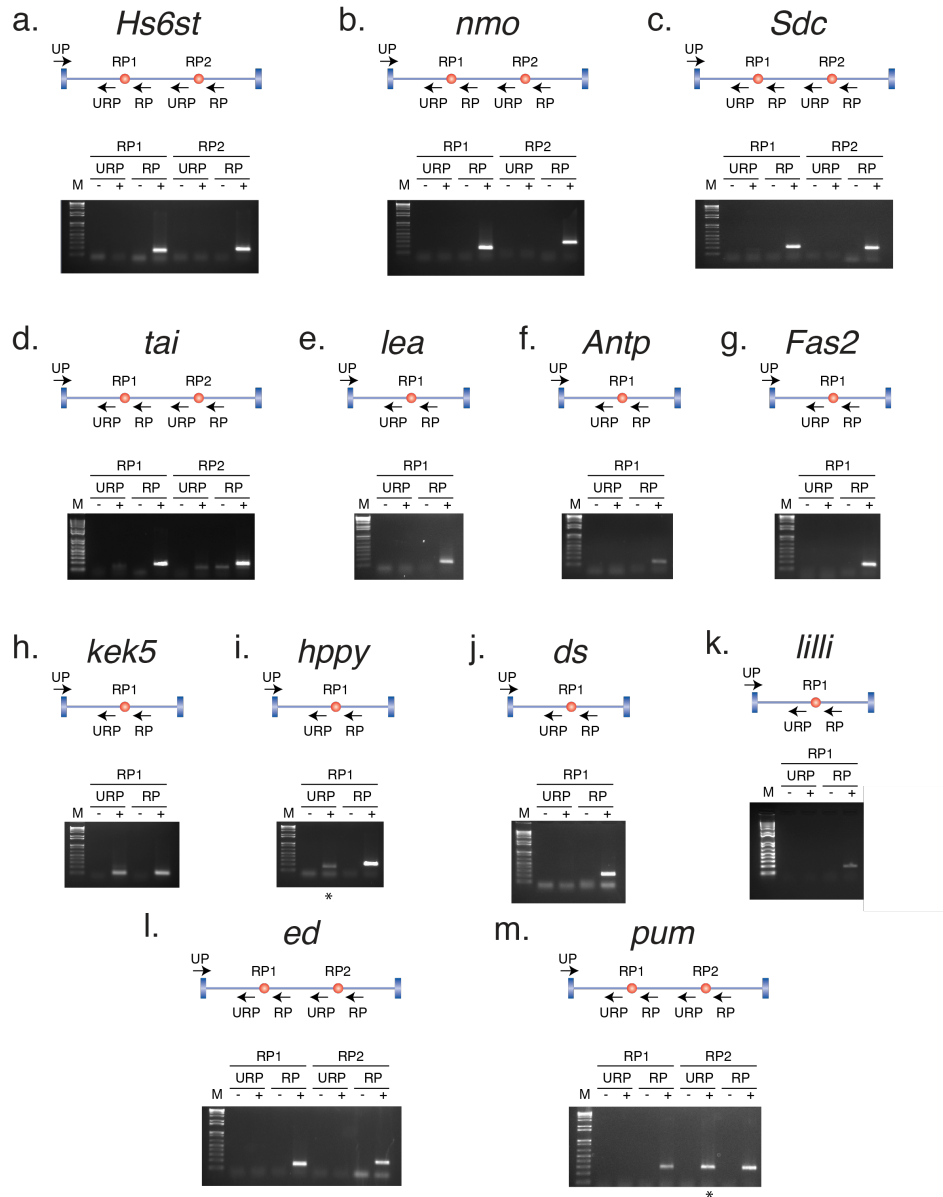
b.



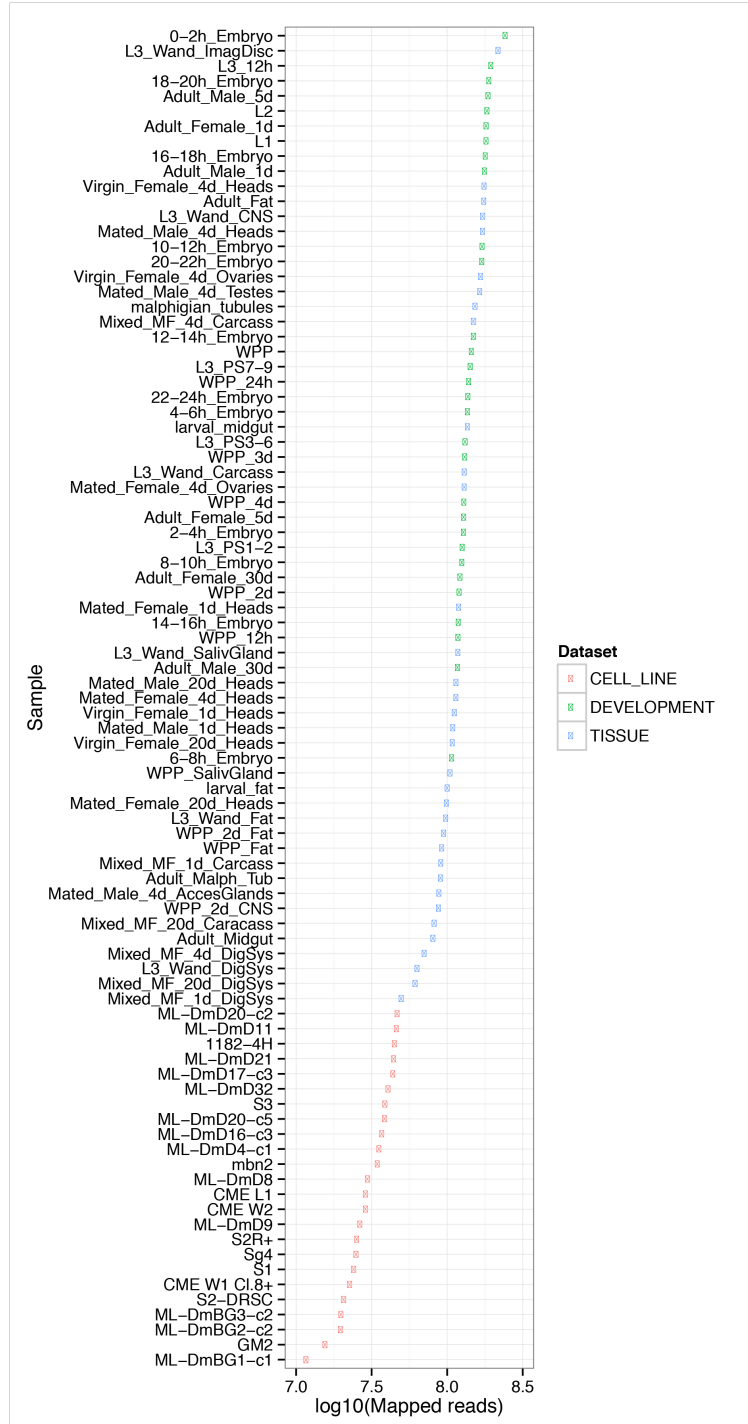
Supplementary Figure 1 | Characteristics of recursive splicing from total RNA-Seq data. **a**, Schematic diagram of nascent pre-mRNA transcripts during co-transcriptional splicing and the corresponding read density that would be observed in total RNA-Seq data. Note the sawtooth pattern created by the 5' to 3' gradient of RNA-Seq read density from the exon to the downstream ratchet point and splice site. **b**, Example of total RNA-Seq data for the *Ubx* gene which is known to contain three recursive splice sites. Also shown are the splice junction reads supporting recursive splicing at each site.



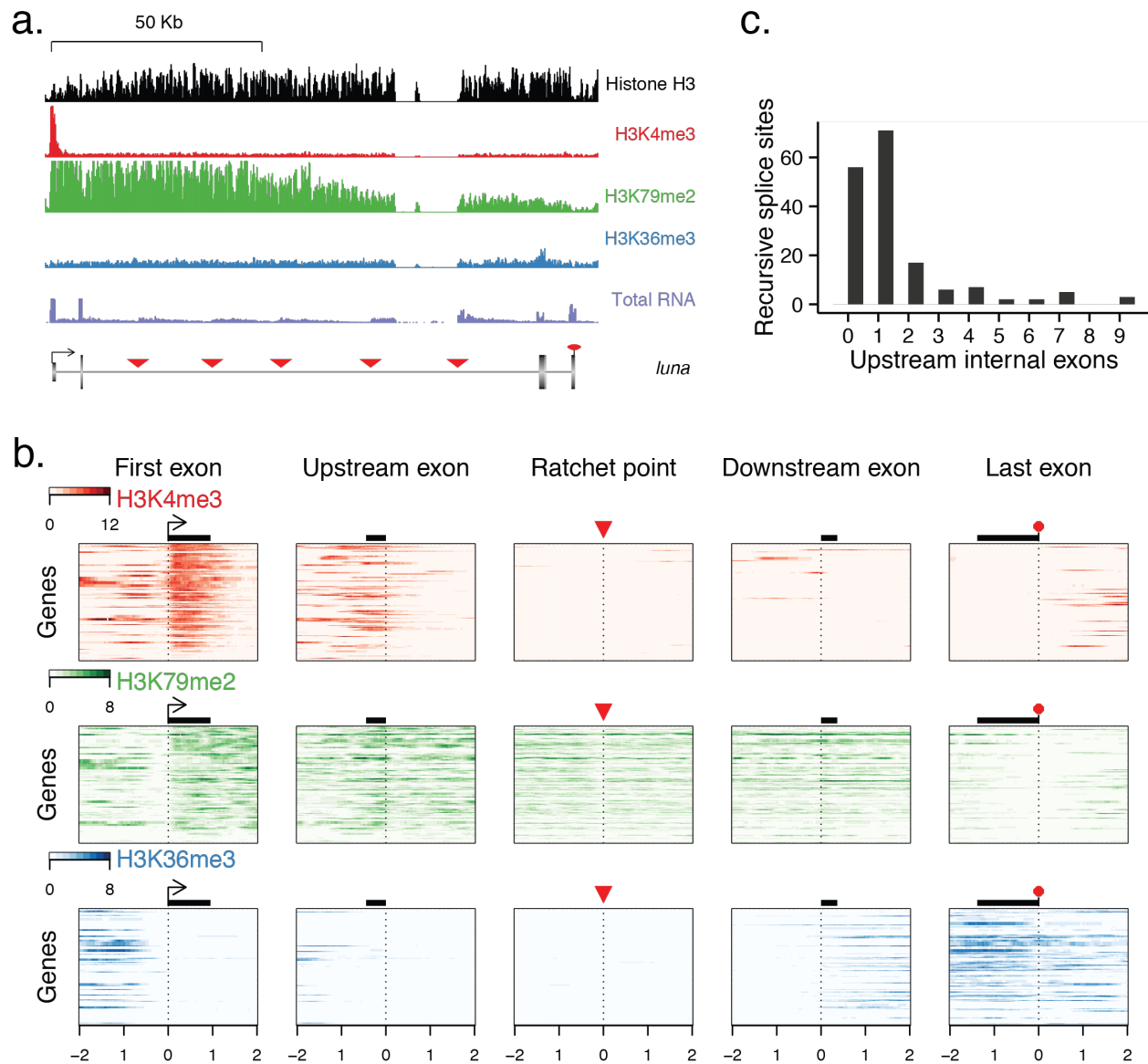
Supplementary Figure 2 | Two approaches for identifying recursive splice sites. a, Identification of recursive splice sites by parsing alignments. RNA-Seq reads were mapped to the genome using TopHat in a manner that allowed for novel splice junctions to be predicted. The alignments were then parsed for splice junction reads where the 5' splice site mapped to an annotated 5' splice site, but the 3' splice site was unannotated. **b, *de novo*** identification of recursive splice sites. A database was generated in which each annotated 5' splice site was spliced to all AG/GT sequences in an intron that did not correspond to an annotated 3' splice site. All RNA-Seq reads were aligned to this database and the alignments parsed to find cases where reads mapped perfectly with at least 3 distinct offsets and at least an 8 nt overhang.



Supplementary Figure 3 | RT-PCR validations of recursive splicing events. RT-PCR validation of ratchet points (red dots) from the indicated genes using primers in the upstream constitutive exon and flanking the putative ratchet points. The RP primers are expected to yield RT-PCR products if the constitutive exon is spliced to the ratchet point. The URP primers, which are upstream of each ratchet point, serve as negative controls. The identity of all RT-PCR products were verified by Sanger sequencing. Though the URP control RT-PCR reactions yielded a product for *hppy* RP1 and *pum* RP2, we were not able to generate sequence from them and therefore consider them to be amplification artifacts.



Supplementary Figure 4 | Number of mapped reads per sample used for gene expression analysis.



Supplementary Figure 5 | Chromatin marks associated with recursive splice sites.

a, Examples of chromatin marks at the *luna* gene locus, which contains 5 recursive splice sites (red triangles) within a single long intron. **b**, Heatmaps show relative ChIP-seq enrichment for H3K4me3 (top, red), H3K79me2 (middle, green), and H3K36me3 (bottom, blue), within 2 kb of the indicated gene features from 171 genes containing at least one ratchet point. Heatmaps are centered around gene features, which include the transcription start site of the first exon (First exon, arrow), the 5' ss of the exon upstream of the recursive splice site (Upstream exon, black rectangle), the ratchet point (red triangle), the 3' ss of the exon downstream of the recursive splice site (downstream

exon, black rectangle), and the poly(A) site of the last exon (last exon, red octagon); the average exon of each gene feature is drawn to scale. Genes are sorted from top to bottom by decreasing expression level. For genes containing more than one ratchet point, the first, upstream, downstream, and last exons are represented multiple times. **c**, Histogram illustrating the number of internal exons upstream of ratchet points based on RefSeq annotations.

Supplementary Tables

Supplementary Table 1. Summary of *D. melanogaster* RNA Sequencing Samples Generated for these studies. For each library, the following information is provided: RNA ID, Biological ID, Biological Sample, Strain (or cell line if appropriate), Data Type (CELL LINE, DEVELOPMENT, ECDYSONE, POPULATION, TISSUE), Total Reads, Uniquely Aligned Reads (from TopHat alignments), % Uniquely Aligned, SRA Accession.

Supplementary Table 2. Summary information for 197 *D. melanogaster* ratchet points. For each ratchet point identified the following information is provided: Ratchet Point ID (chromosome_strand_location), Gene Name, Alias, TopHat aligned reads (total across all samples), de novo aligned reads (total across all samples), de novo mismatches (average frequency of mismatches for all reads across all samples from the de novo alignments), de novo entropy (entropy score for the ratchet points across all samples from the de novo alignments), intron chr (chromosome of the recursive intron), intron start (start position of the recursive intron), intron stop (stop position of the recursive intron), intron strand (strand of the recursive intron), intron chr (chromosome of the recursive intron if more than one exon can be recursively spliced), intron start (start position of the recursive intron if more than one exon can be recursively spliced), intron stop (stop position of the recursive intron if more than one exon can be recursively spliced), intron strand (strand of the recursive intron if more than one exon can be recursively spliced).

Supplementary Table 3. Comparison of ratchet points identified in this study and in Burnette *et al.* All of the predicted ratchet points listed in supplementary Table S1 of Burnette *et al.*⁸ were compared to our experimentally identified ratchet points. To do this, we used BLAT to identify the location of each ratchet point from Burnette *et al.* and tabulated whether or not the ratchet point was identified in our RNA-Seq data. If the ratchet point was not identified, we listed the reason. For each ratchet point the following information is provided: Gene Name, Intron Number, *D. melanogaster* RP Sequence, RP Score (as calculated by Burnette *et al.*), Identified in this study (1 = yes, 0 = no), Reason (exon = ratchet point is an exon, Lack of sawtooth = no compelling sawtooth pattern of RNA-Seq read density observed, missed = not identified in our computational pipeline, no matches = the ratchet point sequence listed from Burnette *et al.* did not match the *D. melanogaster* genome by BLAT).

Supplementary Table 4. Summary of recursive intron lariats identified. For each read aligned to a recursive lariat intron junction, the following information is provided: read ID, strand (of read alignment to the lariat junction), lariat junction ID (build_chr_intronstart_intronstop_strand_RPchr_strand_ratchetpoint:gene:recursive segment_location of 5' end of the 3' segment of the intron included_number of nucleotides included from the 5' end of the intron: e.g., d m 3 _ c h r 3 L _ 1 0 5 1 2 7 1 1 _ 1 0 5 2 7 9 3 0 _ + _ c h r 3 L _ +_10527931:CG32062:segment2_-127_94), alignment start (within the junction), window start (location of 5' end of the 3' segment of the intron included), branchpoint (location of 3' end of the 3' segment of the intron included), Sequence (of aligned read), mismatches, gene, segment (segment of the recursive intron), location (first segment, internal segment, last segment).

Supplementary Table 5. Summary of Total RNA-Seq Data from Related *Drosophila* Species. For each datasets the following information is included: Species, Total reads, Mapped reads (from TopHat alignments), % mapped.

Supplementary Table 6. Summary of ratchet points experimentally validated in other *Drosophila* species. For each species analyzed, the following information is provided: Species, Total leftovers (the number of *D. melanogaster* ratchet point coordinates that were successfully lifted over to the other species genome coordinates), number of ratchet junction reads (the number of reads that mapped to the AG/GT sequence of the lifted over ratchet points), distinct ratchets (the number of distinct ratchet points identified by the mapped reads), % validated (the percent of lifted over ratchet points with at least one mapped read).

Supplementary Table 7. Details of ratchet points experimentally validated in other *Drosophila* species. For each ratchet point in *D. melanogaster*, the number of reads mapping to the orthologous ratchet point in *D. simulans*, *D. sechellia*, *D. yakuba*, *D. pseudoobscura*, or *D. virilis* is indicated. In addition, the total number of reads identified from the other species is indicated in the last column.

Supplementary Table 8. GO analysis of recursively spliced genes. Results of GO analysis of recursively spliced genes using Funcassociate 2.0^(ref. 9).

Supplementary References

- 1 Graveley, B. R. et al. The developmental transcriptome of *Drosophila melanogaster*. *Nature* 471, 473-479, doi:10.1038/nature09715 (2011).
- 2 Brown, J. B. et al. Diversity and dynamics of the *Drosophila* transcriptome. *Nature*, doi:doi:10.1038/nature12962 (2014).
- 3 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105-1111, doi:btp120 [pii] 10.1093/bioinformatics/btp120 (2009).
- 4 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25, doi:10.1186/gb-2009-10-3-r25 (2009).
- 5 Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E. & Fairbrother, W. G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol* 19, 719-721, doi:10.1038/nsmb.2327 (2012).
- 6 Tilgner, H. et al. Nucleosome positioning as a determinant of exon recognition. *Nature Structure & Molecular Biology*, in press (2009).
- 7 Huff, J. T., Plocik, A. M., Guthrie, C. & Yamamoto, K. R. Reciprocal intronic and exonic histone modification regions in humans. *Nat Struct Mol Biol* 17, 1495-1499, doi:10.1038/nsmb.1924 (2010).
- 8 Burnette, J. M., Miyamoto-Sato, E., Schaub, M. A., Conklin, J. & Lopez, A. J. Subdivision of large introns in *Drosophila* by recursive splicing at nonexonic elements. *Genetics* 170, 661-674, doi:10.1534/genetics.104.039701 (2005).
- 9 Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M. & Roth, F. P. Next generation software for functional trend analysis. *Bioinformatics* 25, 3043-3044, doi:10.1093/bioinformatics/btp498 (2009).